Analysing People's Opinion on Brexit from Twitter Data

Abstract

Social media data consists of attitudes, analysis, assessments, and emotions that reflects the same way humans think. Big unstructured data is of the challenges for analysing which are generally found in social media. Some of the social networking sites such as Instagram, Twitter and Facebook give out information about people's sentiments and opinions by providing analysts a platform to collect. By using all these data its easier for analysts to analyse people's opinions. So, by extracting the tweets from Twitter using the developer account and performing certain data cleaning operations it's easy to visualize the opinions that people have on Brexit. The opinions of people may have a good or a bad opinion because not everyone can think the same way. Some people find advantages and some look into the disadvantages. Various visualization methods could be used to visualize and make people understand in an easy way. Here the visualization methods that are used are Word cloud showing the most often used words, a tree map indicating the most used words and the scatter plot showing tweet of the most favourited and retweeted tweet.

1. Dataset

The Dataset is extracted real time from twitter where a user needs to create an account of the Twitter developer account in order to obtain credentials such as API key, API secret, Access token and Access token Secret required to access the Twitter API. After obtaining the credentials one needs to install Twitter library by using the required libraries the user can connect to Twitter API and then download the tweets directly from the Twitter through the Twitter API. There are multiple libraries available and supported by most of the programming languages such as R, Python. Here in this case R was used for data extraction since searchTwitter function allows extraction of tweets of up to 200,000 whereas in Python the Tweepy function allows extraction of only 3200 tweets. In order to extract tweets, one needs to establish a secure connection between the programming language and Twitter. One will be directed to Twitter's authorization screen. Click on Authorize App and note the PIN generated. Go back to the programming language and enter the PIN. Note, this only needs to be done once. Thereby, can successfully access Twitter API and extract tweets. Then finally the tweets are extracted related to Brexit using the searchTwitter function and conditions are given to extract tweets in English language without Retweets.

The Dataset extracted file "tweetsextract.csv" is about 65.7MB and has about 200,000 rows and 17 columns. The attributes are Tweet, id, username, favourited, favourited count, retweet, retweet count, etc.. The data types include Integer, String, Boolean, Date, Float and Double. The aspects involved in the dataset here would be all the 3V's the reason being that the Volume here is 200,000 rows and 17 columns and can be called a big data and is definitely huge a normal person to work and analyse on that huge data. Talking about the Velocity which is the measure of how fast the data is coming is that the data can be extracted real time and then analysed instantly. Twitter is flooded with tweets and many used to analyse the sentiments or the opinions. The Variety of the data in the dataset is there because the reason that not all tweets can be same and if so there might be a few and not all tweets can be same and therefore analysing large distinct tweets is a variety itself.

2. Data Exploration, Processing, Cleaning and/or Integration

The recent Twitter data is extracted from Twitter Developer account based on the term "Brexit" using the R programming language since R supports data extraction of upto 200,000

rows using the searchTwitter function whereas Python allows only extraction of 3200 rows with the Tweepy Function.

The extracted csv file from R programming language is then used in Python for pre processing and cleaning of the required cleaning data set.

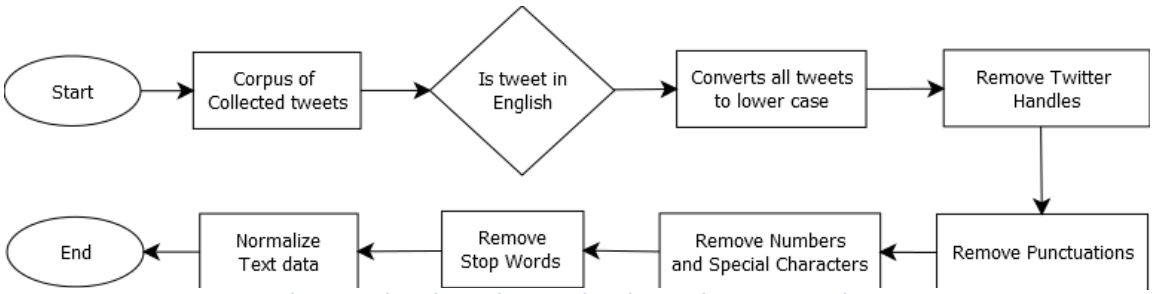

Figure 1 Flowchart of Data Cleaning and Pre-processing

Data pre-processing steps are represented in the form of flowchart as shown in Figure 1. In each step unessential information are removed as discussed in following steps.

Step 1: Removal of Twitter Handles involves removing the "@user", it is a masked term by Twitter to prevent any privacy issues these need to be removed as they don't give any relevant information about the nature of the Tweet. User needs to create a column to store the cleaned and processed tweets. To remove the Twitter handles a regular expression pattern is passed which is "@[]*" so any word starting with "@" is removed.

Step 2: Removal of Numbers, Punctuations and Special Characters since they don't play any significant role in differentiating the types of Tweets. User needs to replace everything else except the characters and hashtags with spaces. The regular expression "[^a-zA-Z#]" means anything except alphabets and '#' thereby removing the punctuations, numbers and special characters.

Step 3: The Small words and the Stop words used in the tweets do not add significant value for the analysis. Words such as "His", "All". So, these kinds of words need to be removed from the Tweet Data. User needs to be cautious while selecting the length of the words to be removed. So, according to politics, the length chosen is 3 so all the words having length less than it will be removed. So, the terms such as "Ha", "Oh" are having no use as they give no relevant information in the tweets they need to be removed.

Step 4: It involves normalizing the Text data. For example, reducing terms like loves, loving, and lovable to their base word, i.e., 'love' is often used in the same context. Normalizing the text helps in reducing the total number of unique words in the tweets data without losing any important/relevant information. Here using Stemming function one can normalize the tweets. But before that tokenization of the tweets needs to be done. Tokens in terms of NLP can be defined as the individual words/terms, and Tokenization can be defined as process which involves splitting a string of text into tokens.

Choosing the attributes to visualize in the case of twitter data is not that of a difficult task the reason being that the twitter data majorly consists of the tweets which has to be cleaned and then broken down and normalized in order to get the word count of each word and analyse the sentiments or the way people react. Then some of the other attributes such as date, retweet count, favourite count, truncated or not are some of the other things used in order to visualize in this assignment.

3. Visualisation

There are 3 visualisation graphs that are created using the twitter dataset.

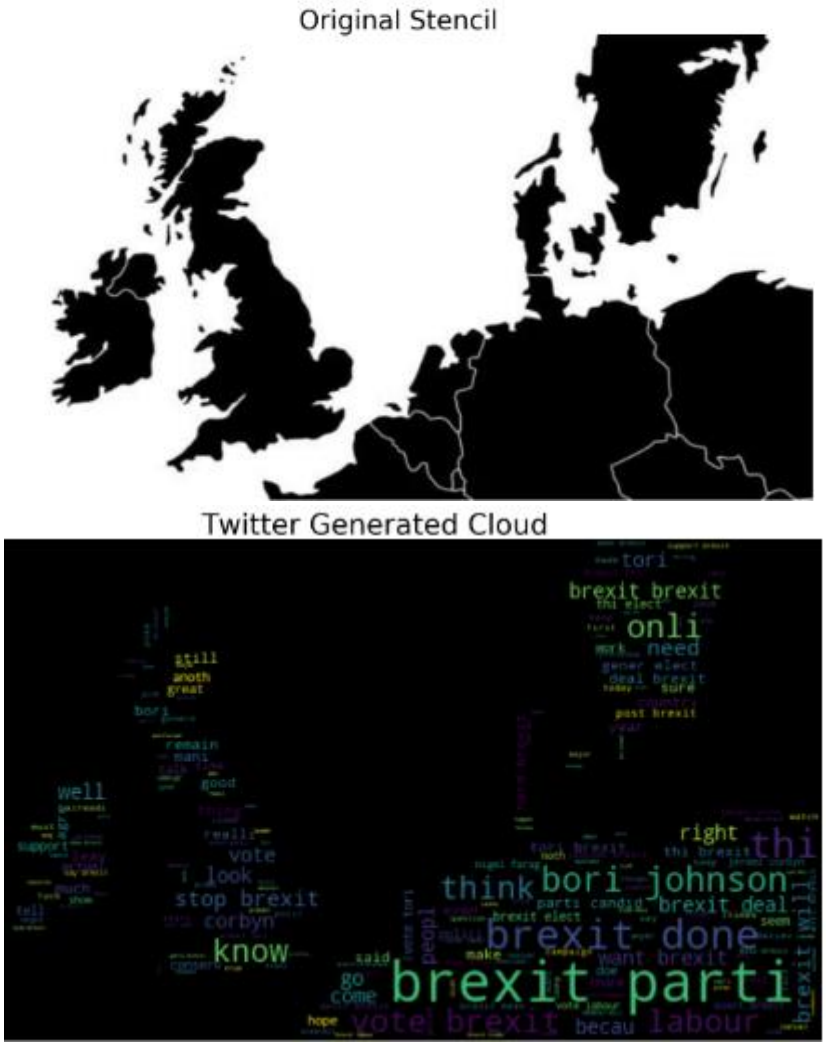

Figure 2 Stencil provided as Input and the Twitter Generated Word Cloud obtained as Output

A word cloud is constructed for pre-processed data set and it represents importance of word in the document. Even though Word Cloud is not that great way of representation the data, it was used mainly because Tweets mostly are in the form of words and word cloud is representation of the words where the size and color of the word represents how frequently a word is being used in the dataset. The shape is masked onto the stencil image as seen in Figure 2, the reason being the image is of the European union, UK and Ireland where all these people are in these countries are eagerly waiting on for Brexit whether its good or Bad. The data type showing is of a string where from the word cloud we can see few words that are huge such as Brexit done, Brexit parti, Bori Johnson, stop Brexit, vote Brexit, etc where each of these have some meaning where all these are trending the reason being that Brexit is still in the verge of happening and there are a lot of tweets by the prime minister Bori Johnson telling that lets get Brexit done. Discussion on parties corresponding to the political parties and political leaders such as Bori Johnson and Tory where some support brexit and some are against Brexit and some would want to stop Brexit from happening whereas some would benefit from Brexit. Basically, it's all Opinions and sentiments people are expressing in the form of tweets. All these things could be analysed from this Word cloud. Here to generate the word cloud Python programming language was used and some of the libraries used are numpy, re, pandas, matplotlib, etc.

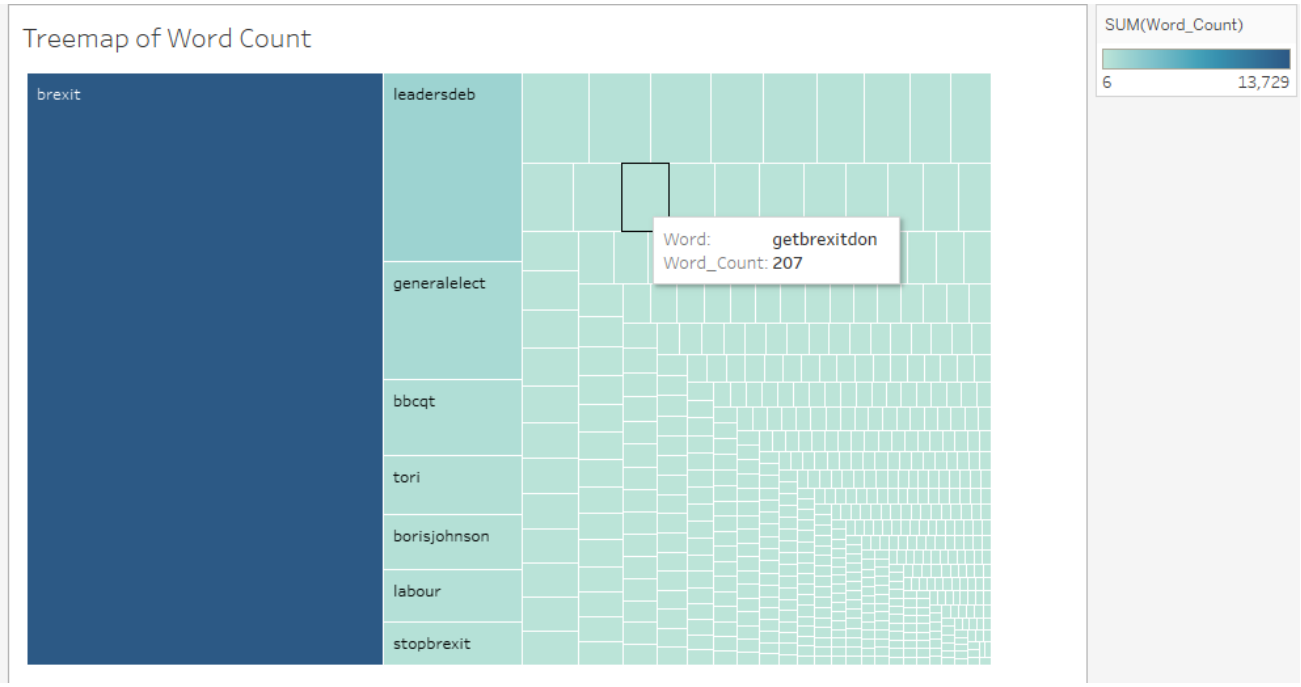

Figure 3 Treemap of the Word count of hashtag

The treemap is constructed after the pre-processing and the data cleaning has been performed where the sentence is broken down for capturing the count of the hastag that has appeared in the tweets from the dataset having 200,000 tweets. All the tweets has gone through the steps that is shown in the Figure 1. From Figure 3 the brexit hastag has the most of the count having 13,729 which can be seen on the right side legend which indicates the color and the figure where the dark blue color that is the Brexit hastag has the highest count and as color of blue diminishes the hastag count decreases. In the legend we can see that the hastag count is ranging from 6 to 13,729. Similarly the bigger the size of the rectangle in the figure 3, the bigger will be the hatag count. The interactivity that can be seen from figure 3 is that the treemap is interactive where each of the rectangular boxes in the treemap highlights the count and the hastag. One more thing that could be analyzed from the dataset that was extracted is that the more people prefer stopping of brexit rather than getting brexit done. This can be justified from the hastag extracted and shown in the figure 4

Word	Word_Count
stopbrexit	400
getbrexitdon	207

Figure 4 Comparing the people's opinions

We can see that the hashtag "Stopbrexit" count is 400 whereas "getbrexitdone" count is 207 and from the dataset that was extracted of about 200,000 tweets I could analyse this that not many want Brexit to happen. The tools used is initially Python where the data cleaning and then the tweets is normalized. After normalizing it is then saved in the csv file and then in tableau the treemap is plotted.

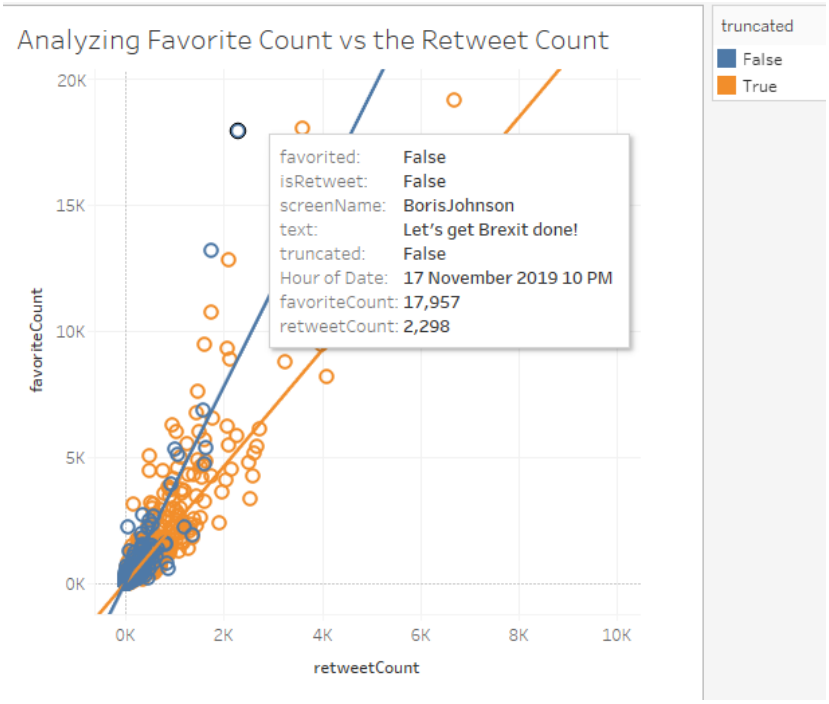

Figure 5 Scattered plot with trend line for analyzing the favourite count and the Retweet count

The scattered plot from Figure 5 gives a very good idea of the dataset as the comparison done is between the retweet count and the favourite count where we can observe from the graph that more the favourite count, the less is the retweet count. Understanding that the retweet count is the number of times people have retweeted or made visible for their followers to have a look at. The favourite count is that the number of people who have liked that following and favourited it. The colours blue and yellow shows that relationship whether or not the tweet is truncated that is if the tweet is shortened when extracting or original tweet having all the 260 words are there. The Blue colour states that the tweet is fully extracted without cutting short the tweet whereas the yellow colour indicates that the tweet has been truncated since tweet is very huge and is having images or figures attached to it. Looking deeply into each of the circle in the scattered plot, it gives details and have answers to a lot of question. The one highlighted in the figure 5 tells that the Tweet was written by Boris Johnson on 17th of November 2019 at 10 PM where the tweet is not truncated and the tweet is not a retweet which means that he has only written the tweet. Then he has not favourited it as well, which can be seen on top if he has favourited that tweet. The main thing is that the Tweet the Prime Minister has tweeted is “Lets get Brexit done” which has around 17,957 favourite count and 2,298 retweet count. Here in Figure 6 the live tweet can be seen in content to the highlighted part in figure 5

Figure 6 The Tweet Boris Johnson Tweeted on 17th November at 10pm

The interactivity that can be observed is that each of the circle in the scatter plot shows the tweet that was tweeted from which person, at what date and time, the retweet count, favourite count, whether truncated or not, whether favourited or not and whether retweeted or not. All of this data can be obtained in a single scatter plot. Also the linear trend line for the truncated and the non-truncated tweets where the best fit line for truncated and non truncated

line can be seen. The tool used is Tableau which is used as a business intelligence tool and here it is giving a lot of details in a single graph which provides an actionable insight and is very fast. Tableau helps a lot in the real-time analysis where in this case Twitter is a social media where 350,000 tweets are tweeted per minute so here the real-time analysis in tableau is very efficient compared to R and Python programming environment.

4. Conclusion

Initially, a twitter developer API account is created. Using the API details and the searchTwitter function 200,000 tweets in English without Retweets on the term Brexit was collected from R programming language and then from the extracted data, the tweets were pre-processed, cleaned and then normalization was performed in Python. Then the individual words and its respective count was extracted and from that the words cloud was created which represents the words where the size and color of the word represents how frequently a word is being used in the dataset. The shape is masked onto the stencil image as seen in Figure 2. Then the hashtag and its respective count data from Python is downloaded in a csv file and then used in Tableau to create a Treemap as shown in Figure 3, the brexit hastag has the most of the count having 13,729 which can be seen on the right side legend which indicates the color and the figure where the dark blue color that is the Brexit hastag has the highest count and as color of blue diminishes the hastag count decreases. Finally again in Tableau the sctter plot shows that each of the circle in the scatter plot shows the tweet that was tweeted from which person, at what date and time, the retweet count, favourite count, whether truncated or not, whether favourited or not and whether retweeted or not. All of this data can be obtained in a single scatter plot.

The one aspect that could have improved is that the use of all the 17th columns in the dataset whereas now only around 10 were effectively used. The challenges that was encountered is the representation of Tweeter extracted data in the form of some animation or some creative graphs as generally the twitter data mostly focuses on the tweets and that can mostly be represented either in a word cloud or a bar plot. The bar plot is the most often ways to visualize the twitter data but here some new visualizations were made and that was possible with Tableau. Python and R doesn't support some innovative and creative graphs.

References

https://developer.twitter.com/
https://developer.twitter.com/en/apps
https://twitter.com/BorisJohnson?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
https://towardsdatascience.com/jupyter-superpower-interactive-visualization-combo-with-python-ffc0adb37b7b